Scriptum Digital
Vol. 7 (2018), pp. 107-110

RESEÑA A TORRUELLA CASAÑAS, JOAN (2017): *LINGÜÍSTICA DE CORPUS: GÉNESIS Y BASES METODOLÓGICAS DE LOS CORPUS (HISTÓRICOS) PARA LA INVESTIGACIÓN EN LINGÜÍSTICA*, (STUDIEN ZUR ROMANISCHEN SPRACHWISSENSCHAFT UND INTERKULTURELLEN KOMMUNIKATION, BAND 116), FRANKFURT AM MAIN: PETER LANG [ISBN: 9783631717189]

MATTHIAS RAAB
Universitat de Barcelona
raab@ub.edu
ORCID-ID: https://orcid.org/0000-0001-6595-0486

Esta obra de Joan Torruella Casañas se corresponde con el volumen 116 de la colección «Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation», editada por Gerd Wotjak y publicada por la editorial Peter Lang. Concretamente, se trata de un manual que pretende, según matiza el propio autor en la «Presentación» (p. 18), «aportar algunas directrices sobre cómo utilizar las *nuevas tecnologías* para la confección de corpus y para el estudio de la lengua a partir de su empleo, mostrando métodos científicos que hagan posible llegar más allá en las investigaciones académicas de lo que se ha podido llegar con el método tradicional». El volumen consta de tres partes, que persiguen sendos objetivos: la primera, titulada «Los corpus y la lingüística de corpus», acota la disciplina lingüística, define los corpus en sí y recoge una tipología de corpus según distintos parámetros; la parte II, «Diseño de la estructura del corpus y post-edición de los textos», versa sobre la preparación, estructuración y elaboración de corpus lingüísticos; en la tercera y última parte, «Bases científicas de la investigación a partir de corpus», se abordan cuestiones metodológicas en torno a la extracción de datos empíricos de los corpus.

La primera parte (pp. 21-60) se organiza en cuatro apartados: «La lingüística de corpus» (pp. 23-29), «Corpus textuales» (pp.31-39), «Parámetros clasificatorios de los corpus» (pp. 41-57) y «Corpus de lectura» (pp. 59-60).

El primer apartado ofrece un marco teórico y un estado de la cuestión amplios sobre la lingüística de corpus, que recoge su historia desde las primeras aproximaciones de lingüistas anglosajones en los años sesenta hasta la actualidad. El autor aborda el acotamiento del área de conocimiento y las distintas opiniones en cuanto a la definición correcta del término *lingüística de corpus*, que puede describirse como «una disciplina, una metodología, una herramienta, una teoría, una praxis o […] una aproximación filosófica» (pp. 25-26). Entre las aproximaciones teóricas, se hallan —entre otras— reflexiones sobre la relación entre la lingüística de corpus con la lingüística computacional, la lingüística generativista o el descriptivismo más tradicional (pp. 27-29).

Fecha de recepción: 29/09/2018
Fecha de admisión: 01/10/2018
ISSN: 2014-640X

En el capítulo 2, siguen abordándose cuestiones teóricas y preliminares a modo introductorio, que, en este caso, pasan de trazar las fronteras de la lingüística de corpus a versar alrededor de su herramienta principal, los corpus lingüísticos. Como en el apartado anterior, se dibuja la historia de los corpus lingüísticos en los últimos siglos y el crecimiento exponencial —tanto de las dimensiones de los corpus en sí como de su explotación en trabajos de carácter científico— desde los años 50 del siglo pasado. Se incide, sobre todo, en los corpus informatizados y la revolución que ha supuesto su uso en la lingüística. Además, se ofrecen, comparan y valoran distintas definiciones de los corpus informáticos.

La tercera sección de la primera parte del manual continúa la labor pionera de Sinclair (1996) y amplía otras propuestas de clasificación tipológica de corpus en castellano, como la del propio Torruella con Llisterri (1999) o la de Cruz (2012: 67-76) —esta última dirigida a expertos en la enseñanza del español—, puesto que se presentan, sin restricciones temáticas, aquellos parámetros que se deben tomar en consideración para definir las características de un corpus: modalidad, temática, época, temporalidad, magnitud, evolución, distribución, número de ediciones, número de lenguas, tipo de edición, muestras y marcaje. Cada uno de estos parámetros y las diversas subcategorías de todos ellos se definen y describen de manera detallada.

Finalmente, el autor insiste en el último apartado de esta primera parte en la importancia de manejar —aparte del corpus informatizado— un corpus de lectura para familiarizarse con el estilo y los contenidos de los textos más destacados del corpus en cuestión.

La segunda parte (pp. 63-233), con diferencia la más larga del manual, seguramente también constituye la más novedosa e interesante, ya que ofrece al lector —entre otras cosas— todas las cuestiones teóricas y prácticas sobre la elaboración de corpus lingüísticos y hace, en el apartado 8, «Preparación de los textos» (pp. 165-233), especial hincapié en el carácter informatizado y algunas técnicas de digitalización de los repertorios y los problemas correspondientes que de ello puedan surgir. El especial interés de dicha parte se debe, probablemente, al desconocimiento profundo sobre metadatos y metalenguajes informáticos del que algunos podamos sufrir.

La parte se divide en cuatro capítulos: «Fases en la construcción de un corpus» (pp. 63-65), que ejerce de parte introductoria y resume la estructura de este segundo bloque del volumen, «Estructura y ejes principales» (pp. 67-128), «Composición del corpus» (pp. 129-163) y «Preparación de textos» (pp. 165-233).

El capítulo sexto —el segundo de esta parte—nos aproxima al objeto de estudio desde la lingüística de variaciones coseriana; y, en concreto desde las variables diacrónica, diatópica y las tipologías textuales. Estos supuestos teóricos se traducen, en el ámbito de los corpus lingüísticos, en los tres ejes fundamentales a la hora de confeccionar un corpus lingüístico: el eje temporal (pp. 72-84), el eje diatópico (pp. 84-93) y el eje tipológico (pp. 93-126). El autor aborda el parámetro temporal a partir de distintas propuestas de periodización intralingüística y sus respectivas limitaciones, y propone posibles organizaciones según características diatópicas basadas en los estudios de dialectología hispánica más destacadas. Tras discutir el eje tipológico de los documentos recogidos en un corpus lingüístico desde la perspectiva de las tradiciones discursivas y las nuevas propuestas de la pragmática histórica, el autor propone una clasificación textual innovadora y exhaustiva según la tipología textual, el tema tratado, el canal, la formalidad y el propósito del texto (p. 111-112). Para los tres ejes, Torruella aporta ejemplos de

clasificación de los corpus diacrónicos más sobresalientes de las lenguas románicas peninsulares: *O Corpus do portugués*, el *Tesouro Medieval Informatizado da Lingua Galega*, el *Corpus del Nuevo Diccionario Histórico del Español*, el *Corpus del español* de Mark Davies, el *Corpus Diacrónico del Español*, el *Corpus de Documentos Españoles anteriores a 1700*, el *Diccionari de textos catalans antics* y el *Corpus Informatitzat del Català Antic*. Para cerrar el capítulo, se anotan algunas deliberaciones sobre el papel de los textos traducidos y la difícil decisión de incluirlos —o no— en el diseño de un corpus lingüístico (pp. 126-128).

Por lo que respecta al capítulo «Composición del corpus», se nos ofrecen explicaciones tanto teóricas como prácticas sobre la confección de un corpus representativo y equilibrado, desde la representatividad cualitativa (pp. 137-138) y cuantitativa (pp. 138-147) hasta la selección de obras y documentos (pp. 149-156). Cobran especial interés filológico los criterios de selección detallados (p. 153) y la inclusión del tema sobre los derechos de autor (*copyright*), que sigue representando una polémica (en este caso más legal que filológica) sin resolver (p. 155). Para los legos y más inexpertos en el etiquetaje informático-digital de los documentos seleccionados, el apartado 7.5, «Filiación de los documentos (Metadatos)» detalla una manera de clasificación según autor, título, fecha, tipo de texto, zona geográfica, etc. (pp. 156-162). El apartado se concluye resumiendo los 10 principios básicos que propone Sinclair (2005: 1-14) para el diseño de cualquier corpus lingüístico (pp. 162-163).

Merece, como hemos apuntado arriba, una lectura pormenorizada y profunda el último capítulo de la segunda parte, que trata, por un lado, sobre la uniformidad de la edición filológica de los textos; y, por otra, sobre la edición filológica digital y cuestiones como la codificación informática o la anotación lingüística de los textos. Aparte de especificar los diferentes tipos de edición textual (facsímil, edición diplomática o paleográfica, edición interpretativa, edición crítica, etc.) (pp. 167-187), el autor puntualiza sobre la codificación de los caracteres, el formato y el nombre de los ficheros (pp. 166-167) y las diferencias entre la edición en papel y en soporte digital (pp. 187-193), y propone algunos criterios sobre el marcaje de las obras mediante distintos procesadores de textos (pp. 193-196) y sistemas de marcación como SGML o XML (pp. 197-198), la estructura de las etiquetas (pp. 198-200), la definición del tipo de documento (pp. 200-205) o los textos TEI (pp. 206-224). Cierra este capítulo el apartado 8.4, «Edición lingüística», sobre la lematización y categorización gramatical de corpus lingüísticos y la estandarización tan necesaria de categorías léxicas, morfológicas, sintácticas, etc. (pp. 224-233).

En el tercer y último bloque del volumen, el autor se adentra en la explotación y el análisis filológico de datos extraídos de corpus lingüísticos. En primer lugar, hallamos una introducción global en métodos de investigación científica, que comprenden la parte introductoria (pp. 235-236) y los capítulos 9, «Elementos base en la investigación científica» (pp. 237-242), y 10, «Método comparativo». Resultan, en este apartado, de gran utilidad, los seis principios de la investigación científica lingüística (pp. 241-242), en que el autor se basa en aquellos formulados por Geeraerts (2013: 43-46). Finalmente, los capítulos 11, «Bases estadísticas en la investigación con corpus» (pp. 247-253), y 12, «El valor de la estadística» (pp. 255-258), concluyen el volumen: se introducen los fundamentos de la investigación cuantitativa aplicada a los corpus lingüísticos; se reseñan los propósitos de los estudios estadísticos; y se describen conceptos imprescindibles, como *variable*, *variante*, *muestreo*, *muestra* o *población*, con sus respectivos subtipos.

En conclusión, el manual reseñado constituye el resultado de la labor científica realizada por Torruella en los últimos años. Nos referimos aquí tanto al ámbito práctico (como responsable del diseño y la confección del corpus *CICA – Corpus Informatitzat del Català Antic*) como teórico —véase una selección de los numerosos estudios sobre lingüística de corpus en el apartado de referencias bibliográficas del volumen (p. 261-279)— o editorial (como codirector de la revista *Scriptum Digital*). Además, aunque se hayan publicado varias obras sobre la lingüística de corpus en el ámbito de la lingüística histórica —véanse, a modo de ejemplo, los volúmenes de Enrique-Arias (2009) o Kabatek en colaboración con De Benito Moreno (2016)—, el volumen representa el primer manual que gira en torno a esta temática.

Por un lado, se reseñan y repasan, de manera profunda y amena, los fundamentos históricos y teóricos de la lingüística de corpus; por otro, y como ya hemos mencionado en líneas anteriores, es de destacar la gran utilidad de la segunda parte —aquella dedicada en parte a cuestiones técnico-informáticas—, dado que a muchos nos puede abrir las puertas a un mundo hasta ahora desconocido. Se trata, en definitiva, de una obra tan innovadora como necesaria que se ha de convertir en lectura obligatoria para todos los estudiosos de la lingüística (no solo) histórica.

REFERENCIAS BIBLIOGRÁFICAS

CRUZ PIÑOL, Mar (2012): *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: Arco/Libros.

ENRIQUE-ARIAS, Andrés (ed.) (2009): *Diacronía de las lenguas iberorrománicas: nuevas aportaciones de la lingüística de corpus.* Madrid/Frankfurt am Main: Iberoamericana/Vervuert.

GERERAERTS, Dirk (2013): «La semántica de corpus cuantitativa como motor de una revolución lingüística basada en las herramientas», en José Francisco Val Álvaro *et al*. (eds.), *De la unidad del lenguaje a la diversidad de las lenguas. Actas del x Congreso Internacional de Lingüística General*. Zaragoza: Universidad de Zaragoza, pp. 35-59.

KABATEK, Johannes y Carlota DE BENITO MORENO (eds.) (2016): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: Walter de Gruyter.

SINCLAIR, John (1996): «Preliminary Recommendations on Corpus Typology», *EAGLES Document EAG-TCWG-CTYP/P*. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

SINCLAIR, John (2005): «Corpus and Text – Basic Principles», in Martin Wynne (ed.), *Developping Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books, pp. 1-16.

TORRUELLA, Joan y Joaquim LLISTERRI (1999): «Diseño de corpus textuales y orales», en José M. Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.), *Filología e Informática: nuevas tecnologías en los estudios filológicos*. Lleida: Milenia/UAB, pp. 45-77.